WIKIPEDIA

# Akaike information criterion

The **Akaike information criterion** (**AIC**) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data.[1][2] Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

AIC is founded on information theory. When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; so some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.

The Akaike information criterion is named after the Japanese statistician Hirotugu Akaike, who formulated it. It now forms the basis of a paradigm for the foundations of statistics and is also widely used for statistical inference.

## Contents

# Definition

Suppose that we have a statistical model of some data. Let $k$ be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following.[3][4]

$$\mathrm{AIC} \,=\, 2k - 2\ln(\hat{L})$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

AIC is founded in information theory. Suppose that the data is generated by some unknown process *f*. We consider two candidate models to represent *f*: $g_1$ and $g_2$. If we knew *f*, then we could find the information lost from using $g_1$ to represent $f$ by calculating the Kullback–Leibler divergence, $D_{\mathrm{KL}}(f \parallel g_1)$; similarly, the information lost from using $g_2$ to represent *f* could be found by calculating $D_{\mathrm{KL}}(f \parallel g_2)$. We would then, generally, choose the candidate model that minimized the information loss.

We cannot choose with certainty, because we do not know *f*. Akaike (1974) showed, however, that we can estimate, via AIC, how much more (or less) information is lost by $g_1$ than by $g_2$. The estimate, though, is only valid asymptotically; if the number of data points is small, then some correction is often necessary (see AICc, below).

Note that AIC tells nothing about the absolute quality of a model, only the quality relative to other models. Thus, if all the candidate models fit poorly, AIC will not give any warning of that. Hence, after selecting a model via AIC, it is usually good practice to validate the absolute quality of the model. Such validation commonly includes checks of the model's residuals (to determine whether the residuals seem like random) and tests of the model's predictions. For more on this topic, see *statistical model validation*.

# How to use AIC in practice

To apply AIC in practice, we start with a set of candidate models, and then find the models' corresponding AIC values. There will almost always be information lost due to using a candidate model to represent the "true model," i.e. the process that generated the data. We wish to select, from among the candidate models, the model that minimizes the information loss. We cannot choose with certainty, but we can minimize the estimated information loss.

Suppose that there are *R* candidate models. Denote the AIC values of those models by $\mathrm{AIC}_1$, $\mathrm{AIC}_2$, $\mathrm{AIC}_3$, ..., $\mathrm{AIC}_R$. Let $\mathrm{AIC}_{\mathrm{min}}$ be the minimum of those values. Then the quantity $\exp((\mathrm{AIC}_{\mathrm{min}} - \mathrm{AIC}_i)/2)$ can be interpreted as being proportional to the probability that the *i*th model minimizes the (estimated) information loss.[5]

As an example, suppose that there are three candidate models, whose AIC values are 100, 102, and 110. Then the second model is exp((100 − 102)/2) = 0.368 times as probable as the first model to minimize the information loss. Similarly, the third model is exp((100 − 110)/2) = 0.007 times as probable as the first model to minimize the information loss.

In this example, we would omit the third model from further consideration. We then have three options: (1) gather more data, in the hope that this will allow clearly distinguishing between the first two models; (2) simply conclude that the data is insufficient to support selecting one model from among the first two; (3) take a weighted average of the first two models, with weights proportional to 1 and 0.368, respectively, and then do statistical inference based on the weighted multimodel.[6]

The quantity exp((AIC$_{\text{min}}$ − AIC$_i$)/2) is known as the *relative likelihood* of model *i*. It is closely related to the likelihood ratio used in the likelihood-ratio test. Indeed, if all the models in the candidate set have the same number of parameters, then using AIC might at first appear to be very similar to using the likelihood-ratio test. There are, however, important distinctions. In particular, the likelihood-ratio test is valid only for nested models, whereas AIC (and AICc) has no such restriction.[7][8]

# Hypothesis testing

Every statistical hypothesis test can be formulated as a comparison of statistical models. Hence, every statistical hypothesis test can be replicated via AIC. Two examples are briefly described in the subsections below. Details for those examples, and many more examples, are given by Sakamoto, Ishiguro & Kitagawa (1986, Part II) and Konishi & Kitagawa (2008, ch. 4).

## Replicating Student's t-test

As an example of a hypothesis test, consider the *t*-test to compare the means of two normally-distributed populations. The input to the *t*-test comprises a random sample from each of the two populations.

To formulate the test as a comparison of models, we construct two different models. The first model models the two populations as having potentially different means and standard deviations. The likelihood function for the first model is thus the product of the likelihoods for two distinct normal distributions; so it has four parameters: $\mu_1$, $\sigma_1$, $\mu_2$, $\sigma_2$. To be explicit, the likelihood function is as follows (denoting the sample sizes by $n_1$ and $n_2$).

$$\mathcal{L}(\mu_1,\sigma_1,\mu_2,\sigma_2) \;=\; \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right) \;\cdot\; \prod_{i=n_1+1}^{n_1+n_2} \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)$$

The second model models the two populations as having the same means but potentially different standard deviations. The likelihood function for the second model thus sets $\mu_1 = \mu_2$ in the above equation; so it has three parameters.

We then maximize the likelihood functions for the two models (in practice, we maximize the log-likelihood functions); after that, it is easy to calculate the AIC values of the models. We next calculate the relative likelihood. For instance, if the second model was only 0.01 times as likely as the first model, then we would omit the second model from further consideration: so we would conclude that the two populations have different means.

The $t$-test assumes that the two populations have identical standard deviations; the test tends to be unreliable if the assumption is false and the sizes of the two samples are very different (Welch's $t$-test would be better). Comparing the means of the populations via AIC, as in the example above, has an advantage by not making such assumptions.

## Comparing categorical data sets

For another example of a hypothesis test, suppose that we have two populations, and each member of each population is in one of two categories—category #1 or category #2. Each population is binomially distributed. We want to know whether the distributions of the two populations are the same. We are given a random sample from each of the two populations.

Let $m$ be the size of the sample from the first population. Let $m_1$ be the number of observations (in the sample) in category #1; so the number of observations in category #2 is $m - m_1$. Similarly, let $n$ be the size of the sample from the second population. Let $n_1$ be the number of observations (in the sample) in category #1.

Let $p$ be the probability that a randomly-chosen member of the first population is in category #1. Hence, the probability that a randomly-chosen member of the first population is in category #2 is $1 - p$. Note that the distribution of the first population has one parameter. Let $q$ be the probability that a randomly-chosen member of the second population is in category #1. Note that the distribution of the second population also has one parameter.

To compare the distributions of the two populations, we construct two different models. The first model models the two populations as having potentially different distributions. The likelihood function for the first model is thus the product of the likelihoods for two distinct binomial distributions; so it has two parameters: $p$, $q$. To be explicit, the likelihood function is as follows.

$$\mathcal{L}(p,q) \;=\; \frac{m!}{m_1!(m-m_1)!}p^{m_1}(1-p)^{m-m_1} \;\cdot\; \frac{n!}{n_1!(n-n_1)!}q^{n_1}(1-q)^{n-n_1}$$

The second model models the two populations as having the same distribution. The likelihood function for the second model thus sets $p = q$ in the above equation; so the second model has one parameter.

We then maximize the likelihood functions for the two models (in practice, we maximize the log-likelihood functions); after that, it is easy to calculate the AIC values of the models. We next calculate the relative likelihood. For instance, if the second model was only 0.01 times as likely as the first model, then we would omit the second model from further consideration: so we would conclude that the two populations have different distributions.

# Foundations of statistics

Statistical inference is generally regarded as comprising hypothesis testing and estimation. Hypothesis testing can be done via AIC, as discussed above. Regarding estimation, there are two types: point estimation and interval estimation. Point estimation can be done within the AIC paradigm: it is provided by maximum likelihood estimation. Interval estimation can also be done within the AIC paradigm: it is provided by likelihood intervals. Hence, statistical inference generally can be done within the AIC paradigm.

The most commonly used paradigms for statistical inference are frequentist inference and Bayesian inference. AIC, though, can be used to do statistical inference without relying on either the frequentist paradigm or the Bayesian paradigm: because AIC can be interpreted without the aid of significance levels or Bayesian priors.[9] In other words, AIC can be used to form a foundation of statistics that is distinct from both frequentism and Bayesianism.[10][11]

# Modification for small sample size

When the sample size is small, there is a substantial probability that AIC will select models that have too many parameters, i.e. that AIC will overfit.[12][13][14] To address such potential overfitting, AICc was developed: AICc is AIC with a correction for small sample sizes.

The formula for AICc depends upon the statistical model. Assuming that the model is univariate, is linear in its parameters, and has normally-distributed residuals (conditional upon regressors), then the formula for AICc is as follows.[15][16]

$$\mathrm{AICc} \;=\; \mathrm{AIC} + \frac{2k^2 + 2k}{n - k - 1}$$

—where $n$ denotes the sample size and $k$ denotes the number of parameters. Thus, AICc is essentially AIC with an extra penalty term for the number of parameters. Note that as $n \to \infty$, the extra penalty term converges to 0, and thus AICc converges to AIC.[17]

If the assumption that the model is univariate and linear with normal residuals does not hold, then the formula for AICc will generally be different from the formula above. For some models, the formula can be difficult to determine. For every model that has AICc available, though, the formula for AICc is given by AIC plus terms that includes both $k$ and $k^2$. In comparison, the formula for AIC includes $k$ but not $k^2$. In other words, AIC is a first-order estimate (of the information loss), whereas AICc is a second-order estimate.[18]

Further discussion of the formula, with examples of other assumptions, is given by Burnham & Anderson (2002, ch. 7) and by Konishi & Kitagawa (2008, ch. 7–8). In particular, with other assumptions, bootstrap estimation of the formula is often feasible.

To summarize, AICc has the advantage of tending to be more accurate than AIC (especially for small samples), but AICc also has the disadvantage of sometimes being much more difficult to compute than AIC. Note that if all the candidate models have the same $k$ and the same formula for AICc, then AICc and AIC will give identical (relative) valuations; hence, there will be no disadvantage in using AIC, instead of AICc. Furthermore, if $n$ is many times larger than $k^2$, then the extra penalty term will be negligible; hence, the disadvantage in using AIC, instead of AICc, will be negligible.

# History

The Akaike information criterion was formulated by the statistician Hirotugu Akaike. It was originally named "an information criterion".[19] It was first announced in English by Akaike at a 1971 symposium; the proceedings of the symposium were published in 1973.[19][20] The 1973 publication, though, was only an informal presentation of the concepts.[21] The first formal publication was a 1974 paper by Akaike.[4] As of October 2014, the 1974 paper had received more than 14,000 citations in the Web of Science: making it the 73rd most-cited research paper of all time.[22]

Nowadays, AIC has become common enough that it is often used without citing Akaike's 1974 paper. Indeed, there are over 150,000 scholarly articles/books that use AIC (as assessed by Google Scholar).[23]

The initial derivation of AIC relied upon some strong assumptions. Takeuchi (1976) showed that the assumptions could be made much weaker. Takeuchi's work, however, was in Japanese and was not widely known outside Japan for many years.

AICc was originally proposed for linear regression (only) by Sugiura (1978). That instigated the work of Hurvich & Tsai (1989), and several further papers by the same authors, which extended the situations in which AICc could be applied.

The first general exposition of the information-theoretic approach was the volume by Burnham & Anderson (2002). It includes an English presentation of the work of Takeuchi. The volume led to far greater use of AIC, and it now has more than 48,000 citations on Google Scholar.

Hirotugu Akaike

Akaike called his approach an "entropy maximization principle", because the approach is founded on the concept of entropy in information theory. Indeed, minimizing AIC in a statistical model is effectively equivalent to maximizing entropy in a thermodynamic system; in other words, the information-theoretic approach in statistics is essentially applying the Second Law of Thermodynamics. As such, AIC has roots in the work of Ludwig Boltzmann on entropy. For more on these issues, see Akaike (1985) and Burnham & Anderson (2002, ch. 2).

# Usage tips

## Counting parameters

A statistical model must account for random errors. A straight line model might be formally described as $y_i = b_0 + b_1x_i + \varepsilon_i$. Here, the $\varepsilon_i$ are the residuals from the straight line fit. If the $\varepsilon_i$ are assumed to be i.i.d. Gaussian (with zero mean), then the model has three parameters: $b_0$, $b_1$, and the variance of the Gaussian distributions. Thus, when calculating the AIC value of this model, we should use $k$=3. More generally, for any least squares model with i.i.d. Gaussian residuals, the variance of the residuals' distributions should be counted as one of the parameters.[24]

As another example, consider a first-order autoregressive model, defined by $x_i = c + \varphi x_{i-1} + \varepsilon_i$, with the $\varepsilon_i$ being i.i.d. Gaussian (with zero mean). For this model, there are three parameters: $c$, $\varphi$, and the variance of the $\varepsilon_i$. More generally, a $p$th-order autoregressive model has $p + 2$ parameters. (If, however, $c$ is not estimated from the data, but instead given in advance, then there are only $p + 1$ parameters.)

## Transforming data

The AIC values of the candidate models must all be computed with the same data set. Sometimes, though, we might want to compare a model of the response variable, $y$, with a model of the logarithm of the response variable, $\log(y)$. More generally, we might want to compare a model of the data with a model of transformed data. Following is an illustration of how to deal with data transforms (adapted from Burnham & Anderson (2002, §2.11.3): "Investigators should be sure that all hypotheses are modeled using the same response variable").

Suppose that we want to compare two models: one with a normal distribution of $y$ and one with a normal distribution of $\log(y)$. We should *not* directly compare the AIC values of the two models. Instead, we should transform the normal cumulative distribution function to first take the logarithm of $y$. To do that, we need to perform the relevant integration by substitution: thus, we need to multiply by the derivative of the (natural) logarithm function, which is $1/y$. Hence, the transformed distribution has the following probability density function:

$$y \mapsto \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right)$$

—which is the probability density function for the log-normal distribution. We then compare the AIC value of the normal model against the AIC value of the log-normal model.

### Software unreliability

Some statistical software will report the value of AIC or the maximum value of the log-likelihood function, but the reported values are not always correct. Typically, any incorrectness is due to a constant in the log-likelihood function being omitted. For example, the log-likelihood function for $n$ independent identical normal distributions is

$$\ln \mathcal{L}(\mu, \sigma) \;=\; -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

—this is the function that is maximized, when obtaining the value of AIC. Some software, however, omits the constant term $(n/2)\ln(2\pi)$, and so reports erroneous values for the log-likelihood maximum—and thus for AIC. Such errors do not matter for AIC-based comparisons, *if* all the models have their residuals as normally-distributed: because then the errors cancel out. In general, however, the constant term needs to be included in the log-likelihood function.[25] Hence, before using software to calculate AIC, it is generally good practice to run some simple tests on the software, to ensure that the function values are correct.

## Comparisons with other model selection methods

The critical difference between AIC and BIC (and their variants) is the asymptotic property under well-specified and misspecified model classes.[26] Their fundamental differences have been well-studied in regression variable selection and autoregression order selection[27] problems. In general, if the goal is prediction, AIC and leave-one-out cross-validations are preferred. If the goal is selection, inference, or interpretation, BIC or leave-many-out cross-validations are preferred. A comprehensive overview of AIC and other popular model selection methods is given by Ding et al. (https://ieeexplore.ieee.org/document/8498082)

### Comparison with BIC

The formula for the Bayesian information criterion (BIC) is similar to the formula for AIC, but with a different penalty for the number of parameters. With AIC the penalty is $2k$, whereas with BIC the penalty is $\ln(n)\,k$.

A comparison of AIC/AICc and BIC is given by Burnham & Anderson (2002, §6.3-6.4), with follow-up remarks by Burnham & Anderson (2004). The authors show that AIC/AICc can be derived in the same Bayesian framework as BIC, just by using different prior probabilities. In the

Bayesian derivation of BIC, though, each candidate model has a prior probability of $1/R$ (where $R$ is the number of candidate models). Additionally, the authors present a few simulation studies that suggest AICc tends to have practical/performance advantages over BIC.

A point made by several researchers is that AIC and BIC are appropriate for different tasks. In particular, BIC is argued to be appropriate for selecting the "true model" (i.e. the process that generated the data) from the set of candidate models, whereas AIC is not appropriate. To be specific, if the "true model" is in the set of candidates, then BIC will select the "true model" with probability 1, as $n \rightarrow \infty$; in contrast, when selection is done via AIC, the probability can be less than 1.[28][29][30] Proponents of AIC argue that this issue is negligible, because the "true model" is virtually never in the candidate set. Indeed, it is a common aphorism in statistics that "all models are wrong"; hence the "true model" (i.e. reality) cannot be in the candidate set.

Another comparison of AIC and BIC is given by Vrieze (2012). Vrieze presents a simulation study—which allows the "true model" to be in the candidate set (unlike with virtually all real data). The simulation study demonstrates, in particular, that AIC sometimes selects a much better model than BIC even when the "true model" is in the candidate set. The reason is that, for finite $n$, BIC can have a substantial risk of selecting a very bad model from the candidate set. This reason can arise even when $n$ is much larger than $k^2$. With AIC, the risk of selecting a very bad model is minimized.

If the "true model" is not in the candidate set, then the most that we can hope to do is select the model that best approximates the "true model". AIC is appropriate for finding the best approximating model, under certain assumptions.[28][29][30] (Those assumptions include, in particular, that the approximating is done with regard to information loss.)

Comparison of AIC and BIC in the context of regression is given by Yang (2005). In regression, AIC is asymptotically optimal for selecting the model with the least mean squared error, under the assumption that the "true model" is not in the candidate set. BIC is not asymptotically optimal under the assumption. Yang additionally shows that the rate at which AIC converges to the optimum is, in a certain sense, the best possible.

## Comparison with cross-validation

Leave-one-out cross-validation is asymptotically equivalent to AIC, for ordinary linear regression models.[31] Asymptotic equivalence to AIC also holds for mixed-effects models.[32]

## Comparison with least squares

Sometimes, each candidate model assumes that the residuals are distributed according to independent identical normal distributions (with zero mean). That gives rise to least squares model fitting.

With least squares fitting, the maximum likelihood estimate for the variance of a model's residuals distributions is the reduced chi-squared statistic, $\hat{\sigma}^2 = \mathrm{RSS}/n$, where $\mathrm{RSS}$ is the residual sum of squares: $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - f(x_i;\hat{\theta}))^2$. Then, the maximum value of a model's log-likelihood function is

$$-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\mathrm{RSS} \;=\; -\frac{n}{2}\ln(\mathrm{RSS}/n) + C$$

—where $C$ is a constant independent of the model, and dependent only on the particular data points, i.e. it does not change if the data does not change.

That gives:[33]

$$\mathrm{AIC} = 2k + n\ln(\sigma^2) - 2C = 2k + n\ln(\mathrm{RSS}) - (n\ln(n) + 2C).$$

Because only differences in AIC are meaningful, the constant $(n\ln(n) + 2C)$ can be ignored, which allows us to conveniently take the following for model comparisons:

$$\Delta\mathrm{AIC} = 2k + n\ln(\mathrm{RSS})$$

Note that if all the models have the same $k$, then selecting the model with minimum AIC is equivalent to selecting the model with minimum RSS—which is the usual objective of model selection based on least squares.

### Comparison with Mallows's $C_p$

Mallows's $C_p$ is equivalent to AIC in the case of (Gaussian) linear regression.[34]

### Bridging the gap between AIC and BIC

A new information criterion, named Bridge Criterion (BC), was developed to bridge the fundamental gap between AIC and BIC.[27] When the data are generated from a finite-dimensional model (within the model class), BIC is known to be consistent, and so is the new criterion. When the underlying dimension is infinity or suitably high with respect to the sample size, AIC is known to be efficient in the sense that its predictive performance is asymptotically equivalent to the best offered by the candidate models; in this case, the new criterion behaves in a similar manner.

## See also

- Bridge Criterion (https://ieeexplore.ieee.org/document/7953690)
- Deviance information criterion
- Focused information criterion
- Hannan–Quinn information criterion
- Maximum likelihood estimation
- Principle of maximum entropy

## Notes

1. McElreath, Richard (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (https://books.google.com/books?id=T3FQDwAAQBAJ&pg=PA189). CRC Press. p. 189. ISBN 978-1-4822-5344-3. "AIC provides a surprisingly simple estimate of the average out-of-sample deviance."
2. Taddy, Matt (2019). Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions (https://books.google.com/books?id=yPOUDwAAQBAJ&pg=PA90). New York: McGraw-Hill. p. 90. ISBN 978-1-260-45277-8. "The AIC is an estimate for OOS deviance."
3. Burnham & Anderson 2002, §2.2
4. Akaike 1974
5. Burnham & Anderson 2002, §2.9.1, §6.4.5
6. Burnham & Anderson 2002

7. Burnham & Anderson 2002, §2.12.4
8. Murtaugh 2014
9. Burnham & Anderson 2002, p. 99
10. Bandyopadhyay & Forster 2011
11. Sakamoto, Ishiguro & Kitagawa 1986
12. McQuarrie & Tsai 1998
13. Claeskens & Hjort 2008, §8.3
14. Giraud 2015, §2.9.1
15. Cavanaugh 1997
16. Burnham & Anderson 2002, §2.4
17. Burnham & Anderson 2004
18. Burnham & Anderson 2002, §7.4
19. Findley & Parzen 1995
20. Akaike 1973
21. deLeeuw 1992
22. Van Noordon R., Maher B., Nuzzo R. (2014), "The top 100 papers (https://www.nature.com/news/the-top-100-papers-1.16224)", Nature, 514.
23. Sources containing both "Akaike" and "AIC" (https://scholar.google.com/scholar?as_vis=0&q=Akaike+AIC&as_sdt=1,5)—at Google Scholar.
24. Burnham & Anderson 2002, p. 63
25. Burnham & Anderson 2002, p. 82
26. Ding, Jie; Tarokh, Vahid; Yang, Yuhong (November 2018). "Model Selection Techniques: An Overview" (https://resolver.caltech.edu/CaltechAUTHORS:20181128-150927005). IEEE Signal Processing Magazine. **35** (6): 16–34. arXiv:1810.09583 (https://arxiv.org/abs/1810.09583). Bibcode:2018ISPM...35...16D (https://ui.adsabs.harvard.edu/abs/2018ISPM...35...16D). doi:10.1109/MSP.2018.2867638 (https://doi.org/10.1109%2FMSP.2018.2867638). ISSN 1053-5888 (https://www.worldcat.org/issn/1053-5888). S2CID 53035396 (https://api.semanticscholar.org/CorpusID:53035396).
27. Ding, J.; Tarokh, V.; Yang, Y. (June 2018). "Bridging AIC and BIC: A New Criterion for Autoregression" (https://ieeexplore.ieee.org/document/7953690). IEEE Transactions on Information Theory. **64** (6): 4024–4043. arXiv:1508.02473 (https://arxiv.org/abs/1508.02473). doi:10.1109/TIT.2017.2717599 (https://doi.org/10.1109%2FTIT.2017.2717599). ISSN 1557-9654 (https://www.worldcat.org/issn/1557-9654). S2CID 5189440 (https://api.semanticscholar.org/CorpusID:5189440).
28. Burnham & Anderson 2002, §6.3-6.4
29. Vrieze 2012
30. Aho, Derryberry & Peterson 2014
31. Stone 1977
32. Fang 2011
33. Burnham & Anderson 2002, p. 63
34. Boisbunon et al. 2014

# References

- Aho, K.; Derryberry, D.; Peterson, T. (2014), "Model selection for ecologists: the worldviews of AIC and BIC", Ecology, **95** (3): 631–636, doi:10.1890/13-1452.1 (https://doi.org/10.1890%2F13-1452.1), PMID 24804445 (https://pubmed.ncbi.nlm.nih.gov/24804445).
- Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), Breakthroughs in Statistics, **I**, Springer-Verlag, pp. 610–624.
- Akaike, H. (1974), "A new look at the statistical model identification", IEEE Transactions on Automatic Control, **19** (6): 716–723, Bibcode:1974ITAC...19..716A (https://ui.adsabs.harvard.edu/abs/1974ITAC...19..716A), doi:10.1109/TAC.1974.1100705 (https://doi.org/10.1109%2FTAC.1974.1100705), MR 0423716 (https://www.ams.org/mathscinet-getitem?mr=0423716).

- Akaike, H. (1985), "Prediction and entropy", in Atkinson, A. C.; Fienberg, S. E. (eds.), A Celebration of Statistics, Springer, pp. 1–24.
- Bandyopadhyay, P. S.; Forster, M. R., eds. (2011), Philosophy of Statistics, North-Holland Publishing.
- Boisbunon, A.; Canu, S.; Fourdrinier, D.; Strawderman, W.; Wells, M. T. (2014), "Akaike's Information Criterion, $C_p$ and estimators of loss for elliptically symmetric distributions", International Statistical Review, **82** (3): 422–439, doi:10.1111/insr.12052 (https://doi.org/10.1111%2Finsr.12052).
- Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference: A practical information-theoretic approach (2nd ed.), Springer-Verlag.
- Burnham, K. P.; Anderson, D. R. (2004), "Multimodel inference: understanding AIC and BIC in Model Selection" (http://www.sortie-nd.org/lme/Statistical%20Papers/Burnham_and_Anderson_2004_Multimodel_Inference.pdf) (PDF), Sociological Methods & Research, **33**: 261–304, doi:10.1177/0049124104268644 (https://doi.org/10.1177%2F0049124104268644), S2CID 121861644 (https://api.semanticscholar.org/CorpusID:121861644).
- Cavanaugh, J. E. (1997), "Unifying the derivations of the Akaike and corrected Akaike information criteria", Statistics & Probability Letters, **31** (2): 201–208, doi:10.1016/s0167-7152(96)00128-9 (https://doi.org/10.1016%2Fs0167-7152%2896%2900128-9).
- Claeskens, G.; Hjort, N. L. (2008), Model Selection and Model Averaging, Cambridge University Press. [Note: the AIC defined by Claeskens & Hjort is the negative of the standard definition—as originally given by Akaike and followed by other authors.]
- deLeeuw, J. (1992), "Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle" (https://web.archive.org/web/20160108225147/http://gifi.stat.ucla.edu/janspubs/1990/chapters/deleeuw_C_90c.pdf) (PDF), in Kotz, S.; Johnson, N. L. (eds.), Breakthroughs in Statistics I, Springer, pp. 599–609, archived from the original (http://gifi.stat.ucla.edu/janspubs/1990/chapters/deleeuw_C_90c.pdf) (PDF) on 2016-01-08, retrieved 2014-11-27.
- Fang, Yixin (2011), "Asymptotic equivalence between cross-validations and Akaike Information Criteria in mixed-effects models" (http://www.jds-online.com/file_download/278/JDS-652a.pdf) (PDF), Journal of Data Science, **9**: 15–21.
- Findley, D. F.; Parzen, E. (1995), "A conversation with Hirotugu Akaike", Statistical Science, **10**: 104–117, doi:10.1214/ss/1177010133 (https://doi.org/10.1214%2Fss%2F1177010133).
- Giraud, C. (2015), Introduction to High-Dimensional Statistics, CRC Press.
- Hurvich, C. M.; Tsai, C.-L. (1989), "Regression and time series model selection in small samples", Biometrika, **76** (2): 297–307, doi:10.1093/biomet/76.2.297 (https://doi.org/10.1093%2Fbiomet%2F76.2.297).
- Konishi, S.; Kitagawa, G. (2008), Information Criteria and Statistical Modeling, Springer.
- McQuarrie, A. D. R.; Tsai, C.-L. (1998), Regression and Time Series Model Selection, World Scientific.
- Murtaugh, P. A. (2014), "In defense of P values" (https://zenodo.org/record/894459), Ecology, **95** (3): 611–617, doi:10.1890/13-0590.1 (https://doi.org/10.1890%2F13-0590.1), PMID 24804441 (https://pubmed.ncbi.nlm.nih.gov/24804441).
- Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. (1986), Akaike Information Criterion Statistics, D. Reidel.
- Stone, M. (1977), "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion", Journal of the Royal Statistical Society, Series B, **39** (1): 44–47, doi:10.1111/j.2517-6161.1977.tb01603.x (https://doi.org/10.1111%2Fj.2517-6161.1977.tb01603.x), JSTOR 2984877 (https://www.jstor.org/stable/2984877).
- Sugiura, N. (1978), "Further analysis of the data by Akaike's information criterion and the finite corrections", Communications in Statistics - Theory and Methods, **7**: 13–26, doi:10.1080/03610927808827599 (https://doi.org/10.1080%2F03610927808827599).

- Takeuchi, K. (1976), " " [Distribution of informational statistics and a criterion of model fitting], Suri Kagaku [Mathematical Sciences] (in Japanese), **153**: 12–18, ISSN 0386-2240 (https://www.worldcat.org/issn/0386-2240).
- Vrieze, S. I. (2012), "Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)", Psychological Methods, **17** (2): 228–243, doi:10.1037/a0027127 (https://doi.org/10.1037%2Fa0027127), PMC 3366160 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3366160), PMID 22309957 (https://pubmed.ncbi.nlm.nih.gov/22309957).
- Yang, Y. (2005), "Can the strengths of AIC and BIC be shared?", Biometrika, **92**: 937–950, doi:10.1093/biomet/92.4.937 (https://doi.org/10.1093%2Fbiomet%2F92.4.937).

# Further reading

- Akaike, H. (21 December 1981), "This Week's Citation Classic" (http://www.garfield.library.upenn.edu/classics1981/A1981MS54100001.pdf) (PDF), Current Contents Engineering, Technology, and Applied Sciences, **12** (51): 42 [Hirotogu Akaike comments on how he arrived at AIC]
- Anderson, D. R. (2008), Model Based Inference in the Life Sciences, Springer
- Arnold, T. W. (2010), "Uninformative parameters and model selection using Akaike's Information Criterion", Journal of Wildlife Management, **74** (6): 1175–1178, doi:10.1111/j.1937-2817.2010.tb01236.x (https://doi.org/10.1111%2Fj.1937-2817.2010.tb01236.x)
- Burnham, K. P.; Anderson, D. R.; Huyvaert, K. P. (2011), "AIC model selection and multimodel inference in behavioral ecology" (https://web.archive.org/web/20170809055834/http://wolfweb.unr.edu/~ldyer/classes/396/burnham2011.pdf) (PDF), Behavioral Ecology and Sociobiology, **65**: 23–35, doi:10.1007/s00265-010-1029-6 (https://doi.org/10.1007%2Fs00265-010-1029-6), S2CID 3354490 (https://api.semanticscholar.org/CorpusID:3354490), archived from the original (https://wolfweb.unr.edu/~ldyer/classes/396/burnham2011.pdf) (PDF) on 2017-08-09, retrieved 2018-05-04
- Cavanaugh, J. E.; Neath, A. A. (2019), "The Akaike information criterion", WIREs Computational Statistics, **11** (3): e1460, doi:10.1002/wics.1460 (https://doi.org/10.1002%2Fwics.1460)
- Ing, C.-K.; Wei, C.-Z. (2005), "Order selection for same-realization predictions in autoregressive processes", Annals of Statistics, **33** (5): 2423–2474, doi:10.1214/009053605000000525 (https://doi.org/10.1214%2F009053605000000525)
- Ko, V.; Hjort, N. L. (2019), "Copula information criterion for model selection with two-stage maximum likelihood estimation" (http://urn.nb.no/URN:NBN:no-77993), Econometrics and Statistics, **12**: 167–180, doi:10.1016/j.ecosta.2019.01.001 (https://doi.org/10.1016%2Fj.ecosta.2019.01.001), hdl:10852/74878 (https://hdl.handle.net/10852%2F74878)
- Larski, S. (2012), The Problem of Model Selection and Scientific Realism (http://etheses.lse.ac.uk/615/1/StanislavLarski_Problem_Model_Selection.pdf) (PDF) (Thesis), London School of Economics
- Pan, W. (2001), "Akaike's Information Criterion in generalized estimating equations", Biometrics, **57** (1): 120–125, doi:10.1111/j.0006-341X.2001.00120.x (https://doi.org/10.1111%2Fj.0006-341X.2001.00120.x), PMID 11252586 (https://pubmed.ncbi.nlm.nih.gov/11252586), S2CID 7862441 (https://api.semanticscholar.org/CorpusID:7862441)

- Parzen, E.; Tanabe, K.; Kitagawa, G., eds. (1998), Selected Papers of Hirotugu Akaike, Springer Series in Statistics, Springer, doi:10.1007/978-1-4612-1694-0 (https://doi.org/10.1007%2F978-1-4612-1694-0), ISBN 978-1-4612-7248-9
- Saefken, B.; Kneib, T.; van Waveren, C.-S.; Greven, S. (2014), "A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models", Electronic Journal of Statistics, **8**: 201–225, doi:10.1214/14-EJS881 (https://doi.org/10.1214%2F14-EJS881)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Akaike_information_criterion&oldid=1033844176"

**This page was last edited on 16 July 2021, at 06:22 (UTC).**